CBS

Colegio Bautista Shalom

Estadística III

Sexto PAE

Segundo Bimestre

Contenidos

ANÁLISIS ESTADÍSTICO

- ✓ COMPRENSIÓN DE LOS DATOS.
  - ELEMENTOS Y COMPETENCIAS EN LA LECTURA DE GRÁFICOS ESTADÍSTICOS.
  - NIVELES DE COMPRENSIÓN DE LOS GRÁFICOS ESTADÍSTICOS.
  - ERRORES EN LA LECTURA Y CONSTRUCCIÓN DE LOS GRÁFICOS.
- ✓ GRÁFICOS ENGAÑOSOS.
  - DIBUJOS.
- ✓ MEDIDAS Y RESÚMENES.
  - MEDIDAS DE POSICIÓN O LOCALIZACIÓN.
  - EL PROMEDIO DE LA MEDIA ARITMÉTICA.
  - MEDIA POBLACIONAL.
  - MEDIA DE DATOS AGRUPADOS.
  - LA MEDIANA MUESTRAL.
  - MEDIANA POBLACIONAL.
  - LA MEDIA α-PODADA.
  - LA MODA.
- ✓ CUARTILES Y OTROS PERCENTILES.
  - CINCO NÚMEROS RESÚMENES.
- ✓ MEDIDAS DE DISPERSIÓN O VARIABILIDAD.
  - RANGO MUESTRAL.
  - DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL.
  - VARIANZA Y DESVIACIÓN ESTÁNDAR POBLACIONAL.

NOTA: conforme avances en tu aprendizaje tu catedrático(a) te indicará la actividad o ejercicio a realizar vaya avanzando con el contenido del presente curso, te indicará la actividad o ejercicio a realizar (como lo considere).

ANÁLISIS ESTADÍSTICO

El análisis estadístico es un componente del análisis de datos.

El análisis estadístico puede ser dividido en cinco pasos discretos, de la siguiente manera:

1. Describir la naturaleza de los datos a ser analizados.
2. Explorar la relación de los datos con la población subyacente.
3. Crear un modelo para resumir la comprensión de cómo los datos se relacionan con la población subyacente.
4. Probar (o refutar) la validez del modelo.
5. Emplear el análisis predictivo para ejecutar escenarios que ayudarán a orientar las acciones futuras.

El objetivo del análisis estadístico es identificar tendencias. Un negocio de venta al por menor, por ejemplo, podría utilizar el análisis estadístico para encontrar patrones en los datos no estructurados y semi-estructurados de los clientes que se puedan utilizar para crear una experiencia para el cliente más positiva y aumentar las ventas.

COMPRENSIÓN DE LOS DATOS

El lenguaje gráfico tiene un papel esencial en la organización, descripción y análisis de datos, al ser un instrumento de *transnumeración.* Esta es una de las formas básicas de razonamiento estadístico definidas por Wild y Pfannkuch (1999), que consiste en obtener una nueva información, al cambiar de un sistema de representación a otro. Por ejemplo, al pasar de un listado de datos a un histograma, puedes percibir el valor de la moda, que antes no era visible en los datos brutos.

La construcción e interpretación de gráficos estadísticos es también parte importante de la cultura estadística a la que cada vez se dedica más atención y que Gal (2002, pg. 2) define como la unión de dos competencias relacionadas*:*

a) *Interpretar y evaluar críticamente la información estadística, los argumentos apoyados en datos o los fenómenos estocásticos que las personas pueden encontrar en diversos contextos, incluyendo los medios de comunicación, pero no limitándose a ellos; y,*
b) *Discutir o comunicar sus opiniones respecto a tales informaciones estadísticas cuando sea relevante.* (Gal, 2002, pp. 2-3).

A pesar de esta importancia de los gráficos estadísticos, la investigación en didáctica de la matemática nos alerta que la competencia relacionada con el lenguaje de las gráficas estadísticas no se alcanza en la educación obligatoria (Cazorla, 2002) ni tampoco en la preparación de los futuros profesores de Educación Primaria (Espinel, 2007).

En lo que sigue ampliamos un trabajo de Surrey sobre comprensión de gráficos (Batanero, Arteaga y Díaz, en prensa) para informar sobre las investigaciones relacionadas con las competencias y niveles de comprensión de gráficos estadísticos, así como errores frecuentes relacionados con los mismos.

ELEMENTOS Y COMPETENCIAS EN LA LECTURA DE GRÁFICOS ESTADÍSTICOS

Un primer punto investigado por diversos autores es la competencia en la lectura de gráficos. Encontramos gráficos en la prensa diaria, en Internet y también en textos de materias como las ciencias sociales. Sería por tanto necesario que una persona culta fuese capaz de comprender la información expresada en los mismos, aunque esta competencia no es sencilla. Cuando se pide a un estudiante interpretar un gráfico, el estudiante debe realizar la traducción entre lo representado en el gráfico y la realidad. Pero esta traducción requiere conocimientos tanto sobre la realidad representada, como sobre los convenios de construcción del gráfico que a veces el estudiante no posee.

Un gráfico queda determinado por los siguientes elementos (Curcio, 1987):

- ✓ Las *palabras* que aparecen en el gráfico, como el título del gráfico, las etiquetas de los ejes y de las escalas, y que proporcionan las claves necesarias para comprender el contexto, las variables y las relaciones expresadas en el gráfico.
- ✓ El *contenido matemático* subyacente en el gráfico. Por ejemplo, los conjuntos numéricos empleados y otros conceptos matemáticos implícitos en el gráfico que el estudiante ha de dominar para interpretarlo, como los de área en un gráfico de sectores, longitud en un gráfico de líneas o sistema de coordenadas cartesianas en un diagrama de dispersión.

- ✓ Los *convenios específicos* que se usan en cada tipo de gráfico y que se deben conocer para poder realizar una lectura o construcción correcta. Por ejemplo, el alumno ha de conocer que, en un diagrama de sectores, la amplitud del sector es proporcional a la frecuencia. En diagrama de dispersión, cada punto representa un caso y las coordenadas del punto los valores de las dos variables representadas. En algunos gráficos estadísticos estos convenios no son sencillos, como ocurre en el gráfico de la caja, que es muy difícil de interpretar si no se estudia la forma en que se construye e interpreta.

  Partiendo del análisis anterior, Friel, Curcio y Bright (2001) identifican los siguientes elementos estructurales de un gráfico estadístico:
- ✓ El *título* y las *etiquetas* indican el contenido contextual del gráfico y cuáles son las variables en él representadas. Será importante incluir un título y etiquetas no ambiguas.
- ✓ El *marco* del gráfico, que incluye los ejes, escalas, y marcas de referencia en cada eje. Dicho marco proporciona información sobre las unidades de medida de las magnitudes representadas. Puede haber diferentes tipos de marcos y sistemas de coordenadas (lineales, cartesianas bidimensionales o multidimensionales, polares).

Los *especificadores* del gráfico son los elementos usados para representar los datos, como los rectángulos (en el histograma) o los puntos (en el diagrama de dispersión). Los autores nos alertan de que no todos los especificadores son igualmente sencillos de comprender, sugiriendo el siguiente orden de dificultad: Posición en una escala homogénea (gráficos de línea, de barras, de puntos, algunos pictogramas e histogramas); posición en una escala no homogénea (gráficos polares, gráficos bivariantes); longitud (gráficos poligonales o estrellados sin ejes de referencia, árboles), ángulo o pendiente (gráfico de sectores, discos), área (círculos, pictogramas), volumen (cubos, algunos mapas estadísticos), color (mapas estadísticos codificados mediante color).

En relación con los anteriores componentes del gráfico, su lectura y construcción, se requieren, según Friel, Curcio y Bright (2001), los siguientes tipos de competencias relacionadas con el lenguaje de los gráficos:

- ✓ Reconocer los elementos estructurales del gráfico (ejes, escalas, etiquetas, elementos específicos) y sus relaciones. Esta competencia se adquiere cuando es posible distinguir cada uno de estos elementos y si cada elemento es o no apropiado en el gráfico particular.
- ✓ Apreciar el impacto de cada uno de estos componentes sobre la presentación de la información en un gráfico (por ejemplo, ser capaz de predecir como cambiaría el gráfico al variar la escala de un eje).
- ✓ Traducir las relaciones reflejadas en el gráfico a los datos que se representan en el mismo y viceversa. Por ejemplo, cuando un diagrama de dispersión es creciente, comprender que la relación representada entre las dos variables es directa.
- ✓ Reconocer cuando un gráfico es más útil que otro, en función del juicio requerido y de los datos representados, es decir, saber elegir el gráfico adecuado al tipo de variable y al tipo de problema.

NIVELES DE COMPRENSIÓN DE LOS GRÁFICOS ESTADÍSTICOS

Además de las competencias anteriores, algunos autores definen niveles de comprensión en la lectura crítica de datos y muestran que no todos los alumnos alcanzan el nivel más alto durante el diversicado. A continuación, resumimos las teorías de diversos autores al respecto.

Bertin (1967) sugiere que la lectura de un gráfico comienza con una *identificación externa,* del tema al que se refiere el gráfico, a través de la comprensión del significado del título y las etiquetas. A continuación se requiere una *identificación interna,* de las dimensiones relevantes de variación en el gráfico: variables representadas y escala. Finalmente se produce una *percepción de la correspondencia* entre los niveles particulares de cada dimensión visual para obtener conclusiones sobre los niveles particulares de cada variable y sus relaciones en la realidad representada.

A partir de estos supuestos define diversos niveles de lectura de un gráfico:

- ✓ *Extracción de datos*, que consiste en poner en relación un elemento de un eje con el de otro eje. Por ejemplo, en un diagrama de barras leer la frecuencia asociada a un valor de la variable o bien en un diagrama de dispersión leer las coordenadas de uno de los puntos.
- ✓ *Extracción de tendencias,* cuando se es capaz de percibir en el gráfico una relación entre dos subconjuntos de datos que pueden ser definidos a priori o visualmente. Por ejemplo, al determinar visualmente la moda de una distribución en un diagrama de barras, se clasifican los datos en subconjuntos (que tienen un mismo valor para la variable) y se comparan entre si estos

subconjuntos para ver cuál tiene mayor frecuencia. Otro ejemplo sería detectar la simetría o asimetría de una distribución a partir de su representación en un histograma.

- ✓ *Análisis de la estructura* de los datos, comparando tendencias o agrupamientos y efectuando predicciones. Por ejemplo, cuando se representa en un diagrama de barras adosadas dos distribuciones y se analizan las diferencias en promedios y dispersión de las mismas.

Otra clasificación en niveles de comprensión de los gráficos, muy similar a la anterior, con un gran impacto en educación estadística se debe a Curcio (1989), que denominó a los tres niveles definidos por Bertin como *leer entre los datos* (lectura literal del gráfico sin interpretar la información contenida en el mismo), *"leer dentro de los datos"* (interpretación e integración de los datos en el gráfico) y "*leer más allá de los datos*" (realizar predicciones e inferencias a partir de los datos sobre informaciones que no se reflejan directamente en el gráfico). Este autor mostró que las principales dificultades aparecen en los dos niveles superiores y que el nivel progresa con la edad de los estudiantes. Friel, Curcio y Bright (2001) amplían la clasificación anterior definiendo un nuevo nivel *leer detrás de los datos* consistente en valorar críticamente el método de recogida de datos, su validez y fiabilidad, así como las posibilidades de extensión de las conclusiones.

ERRORES EN LA LECTURA Y CONSTRUCCIÓN DE LOS GRÁFICOS

El primer paso en su construcción sería elegir un gráfico adecuado, tanto al tipo de variable, como al problema planteado, pero los estudiantes fallan con frecuencia en esta elección.

- ✓ Elegir una escala inadecuada para el objetivo pretendido (por ejemplo, no se cubre todo el campo de variación de la variable representada)
- ✓ Omitir las escalas en alguno de los ejes horizontal o vertical, o en ambos.
- ✓ No especificar el origen de coordenadas.
- ✓ No proporcionar suficientes divisiones en las escalas de los ejes.

Encontramos también investigaciones sobre errores en la lectura y comprensión de gráficos específicos. En el diagrama de barras, al variar la disposición de los datos (por ejemplo, al usar barras horizontales en lugar de verticales) los estudiantes pueden tener errores simples de lectura (Pereira-Mendoza y Mellor, 1990). Lee y Meletiou (2003) nos alertan de cuatro principales categorías de razonamientos erróneos a la hora de construir, interpretar y aplicar los histogramas en diferentes contextos de la vida real:

Percepción de los histogramas como representación de datos aislados, suponiendo que cada rectángulo se refiere a una observación particular y no a un intervalo de valores.

Tendencia a observar el eje vertical y comparar las diferencias en las alturas de las barras cuando comparan la variación de dos histogramas.

Interpretación determinista, sin apreciar que los datos representan un fenómeno aleatorio que podría variar al tomar diferentes muestras de la misma población. Tendencia a interpretar los histogramas como gráficos de dos variables (es decir, como diagramas de dispersión).

Respecto a los gráficos de caja, aunque su estudio se recomienda en la Educación Secudaria Obligatoria, Bakker, Biehler y Konold (2004) indican que esta representación no permite a los estudiantes percibir los valores individuales de los datos y además son muy diferentes a otros gráficos usados por los alumnos ya que están basados en la mediana y cuartiles, conceptos que no son intuitivos para los alumnos.
El ordenador en ocasiones contribuye a empeorar los problemas de los estudiantes. Ben-Zvi y Friedlander (1997) analizan los gráficos producidos por sus alumnos al trabajar con proyectos de análisis de datos con ayuda del ordenador, identificando cuatro categorías:

- ✓ *Uso acrítico:* los estudiantes construyen gráficos rutinariamente aceptando las opciones por defecto del software, aunque no sean adecuadas. Tienen también dificultad en valorar las relaciones sugeridas en sus representaciones gráficas, identificando solo la información obvia, por ejemplo, los máximos.
- ✓ *Uso significativo de una representación:* los estudiantes construyen correctamente un gráfico si se les indica cuál ha de utilizar; también lo pueden justificar en base al tipo de datos o al problema planteado. Son capaces de modificar y transformar la gráfica (por ejemplo, cambiar una opción del software) e interpretan los resultados, pero no son capaces de seleccionar la gráfica más adecuada cuando tienen varias para elegir.

- ✓ *Manejo significativo de representaciones múltiples:* en este caso, los alumnos toman decisiones en la selección de los gráficos más adecuados, toman en consideración cuál es la contribución de éstos a su problema.
- ✓ *Uso creativo:* cuando el alumno elabora un gráfico correcto, no habitual, para presentar y justificar sus ideas.

GRÁFICOS ENGAÑOSOS

A menudo los gráficos que se presentan son engañosos, es decir, no reflejan adecuadamente los resultados o exageran ciertas características de los datos. Veremos algunas situaciones.

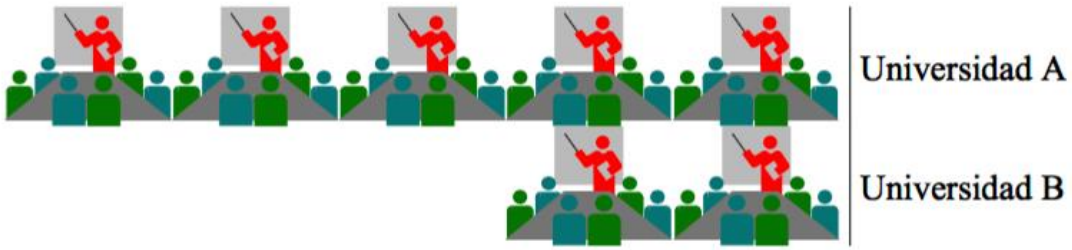
DIBUJOS

En la Figura 1 se representa el número de conferencias organizadas en todos los departamentos de la Universidad A y la Universidad B, en el año 2000.

Cada ícono representa 20 conferencias, por lo tanto, el gráfico informa que en la Universidad A se organizaron aproximadamente 100 conferencias en tanto que en B se organizaron 40.

La información que brinda el gráfico es equivalente a la información numérica.

Figura 1. Número de conferencias organizadas por las Universidades A y B en 2000 (*).

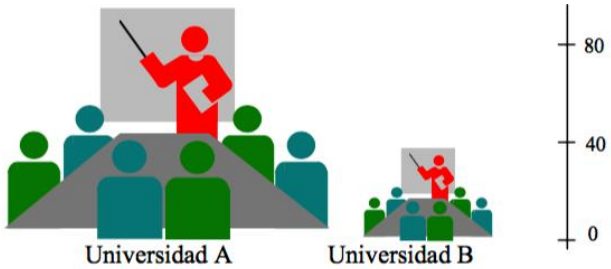


(*) *Cada ícono representa 20 conferencias.*

Cuando la representación se realiza utilizando símbolos que cambian de tamaño, la imagen puede resultar engañosa, tal como ocurre al representar los datos anteriores en la la Figura 2. En esta

Figura, la altura del ícono indica el número de conferencias. La impresión visual es engañosa porque no está claro cuál de las dimensiones de la figura representa la magnitud de la variable. En general, frente a dibujos que no tienen la misma base, tendemos a comparar áreas.

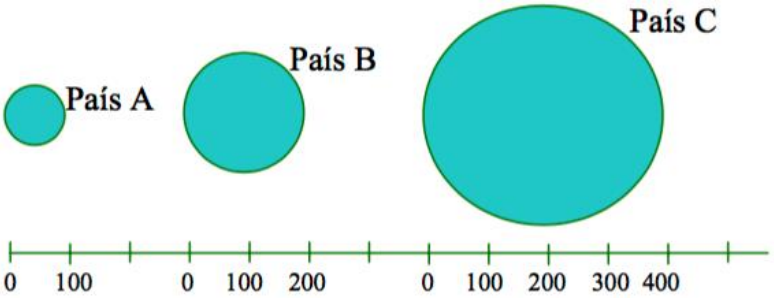
Número de conferencias organizadas en las Universidades A y B en 2000 (*)



(*) *El número de conferencias se representa en la altura del ícono.*

La Figura 3 es otro ejemplo de la misma situación. Como las magnitudes se representan en el diámetro, aun cuando el diámetro de B es el doble que el de A, como el área de B es 4 veces la de A, el gráfico produce una impresión engañosa.

Figura 3. Deuda externa de 3 países (en miles de millones de dólares) (a)



(a) *La deuda se representa en el diámetro.*

El punto clave aquí es que aun cuando el gráfico es correcto, solo será correctamente interpretado por los pocos lectores acostumbrados a leer los detalles de las notas al pie.

MEDIDAS Y RESÚMENES

Aprenderás distintas formas de resumir la distribución muestral o poblacional de una variable NUMÉRICA y finalmente presentaremos un tipo de gráfico que se construye a partir de medidas resúmenes.

Resumir un conjunto de datos es pasar de una visión detallada a una generalización simple e informativa tratando de preservar las características esenciales.

¿Por qué resumir? Para simplificar la comprensión y la comunicación de los datos.

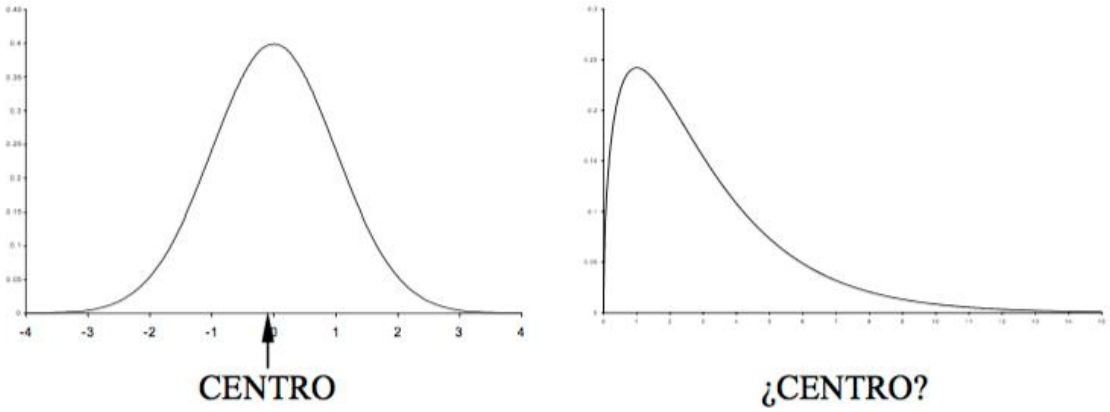
Las medidas resúmenes son útiles para comparar conjuntos de datos cuantitativos y para presentar los resultados de un estudio y se clasifican en dos grupos principales:

Medidas de posición o localización ⇒	describen un valor alrededor del cual se encuentran las observaciones
Medidas de dispersión o escala ⇒	pretenden expresar cuan variable es un conjunto de datos.

MEDIDAS DE POSICIÓN O LOCALIZACIÓN

Un modo de resumir un único conjunto de datos numéricos es a través de un número que debería ser típico para el grupo. No debería ser ni demasiado grande, ni demasiado pequeño y debería estar tan cerca del "centro" de la distribución como sea posible.

Por lo tanto, una medida de posición es un número que pretende indicar dónde se encuentra el centro de la distribución de un conjunto de datos. Pero, ¿dónde se encuentra el "centro" de una distribución?



El centro es fácil de identificar si la distribución es simétrica, pero es difícil si la distribución es asimétrica. Por esta razón, no hay una única medida de posición para resumir una distribución. Si la distribución es simétrica diferentes medidas conducirán a similares resultados. Si la distribución es claramente asimétrica diferentes propuestas apuntarán a distintos conceptos de "centro" y por lo tanto los valores serán diferentes.

A los efectos de resumir los datos debemos preguntarnos:

- ✓ ¿Qué medida resumen es la más apropiado para la distribución que presentan nuestros datos?
- ✓ ¿Qué propuesta permite responder mejor a las preguntas sobre el mundo real que pretendemos responder con estos datos?

EL PROMEDIO DE LA MEDIA ARITMÉTICA

Es la medida de posición más frecuentemente usada. Para calcular la media aritmética o promedio de un conjunto de observaciones se suman todos los valores y se divide por el número total de observaciones.

Definición

Si tenemos una muestra de n observaciones y denotadas por $X_1, X_2, ..., X_n$, definimos la *media muestral* $\overline{X}$ del siguiente modo:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

El símbolo $\sum_{i=1}^{n} X_i$ indica la suma de todos los valores obesrvados de la variable desde el primero ($i = 1$) hasta el último ($i = n$).

Ejemplo.

$X_1 = 10 \qquad X_2 = 14 \qquad X_3 = 12 \qquad X_4 = 11 \qquad X_5 = 12 \qquad X_6 = 13$

$$\overline{X} = \frac{X_1 + X_2 + ... + X_6}{n} = \frac{10+14+12+11+12+13}{6} = \frac{72}{6} = 12$$

MEDIA POBLACIONAL

Si se dispone de la información de una variable X para las N unidades de análisis de la población, es posible calcular la media poblacional a la que denotaremos con la letra griega μ (mu), para distinguirla de la media obtenida en una muestra de n

$$\mu = \frac{X_1 + X_2 + ... + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}.$$

MEDIA DE DATOS AGRUPADOS

Supongamos que se dispone de dos conjuntos de datos en los que se conoce la media y el número de datos de cada uno de ellos $(\overline{X}_1, n_1$ y $\overline{X}_2, n_2)$. Calculamos la media de los $n_1 + n_2$ datos como el *promedio pesado*

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

Ejemplo. Datos sobre niveles de hierro sérico en niños y niñas con fibrosis cística. X = nivel de hierro sérico

	Varones	Mujeres
$\overline{X}$	5.9	6.8
n	13	6

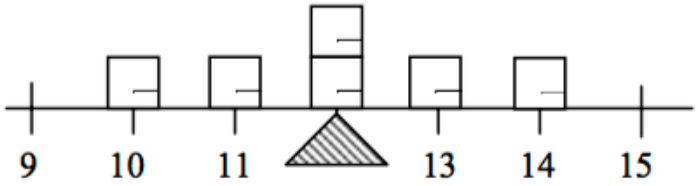
$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} = \frac{13 \cdot 5.9 + 6 \cdot 6.8}{13+6} = \frac{(76.7 + 40.8)}{19} = \frac{117.5}{19} = 6.18$$

El promedio pesado obtenido aquí es igual al que hubiéramos obtenido promediando los datos de los 19 niños.

Características y propiedades de la media:

a) Se usa para datos numéricos.
b) Representa el centro de gravedad o el punto de equilibrio de los datos.

Podemos imaginar a los datos como un sistema físico, en el que cada dato tiene una "masa" unitaria y lo ubicamos sobre una barra en la posición correspondiente a su valor. La media representa la posición en que deberíamos ubicar el punto de apoyo para que el sistema esté en equilibrio.



c) La suma de las distancias de los datos a la media es cero. Esta propiedad está relacionada con el hecho que la media es el centro de gravedad de los datos.

En la tabla siguiente comprobamos esta propiedad para los datos del ejemplo anterior.

X_i	$X_i - \overline{X}$
10	-2
14	2
12	0
111	-1
12	0
13	1
Total =	0

d) Es muy sensible a la presencia de datos atípicos (OUTLIERS).

Modificamos 1 dato en el ejemplo anterior $X_2 = 14 \rightarrow X_2 = 26$ y $\overline{X} = 12 \rightarrow \overline{X} = 14$.

Con solo modificar un dato la media se desplazó tanto, que ya no se encuentra entre la mayoría de los datos. Podemos decir que en este caso la media no es una buena medida de posición de los datos. En consecuencia, la media es una buena medida del centro de la distribución cuando esta es simétrica.

Aunque la media es una medida simple de tendencia central, otras medidas son más informativas y ocasionalmente más apropiadas.

LA MEDIANA MUESTRAL

La mediana es el dato que ocupa la posición central en la muestra ordenada de menor a mayor.

¿Cómo calculamos la mediana de una muestra de n observaciones?

1. Ordenamos los datos de menor a mayor.
2. La mediana es el dato que ocupa la posición $\left(\frac{n+1}{2}\right)$ en la lista ordenada.

Si el número de datos es *impar*, la mediana $\tilde{X}$ es el dato que ocupa la posición central.

Si el número de datos es *par*, la mediana $\tilde{X}$ es el promedio de los dos datos centrales.

Ejemplo:

- *n impar*

$X_1 = 10 \quad X_2 = 14 \quad X_3 = 12 \quad X_4 = 18 \quad X_5 = 11$

Ordenamos los datos:

10 11 12 14 18

La posición de la mediana es $\frac{n+1}{2} = \frac{5+1}{2} = 3$ (tercer dato), es decir $\tilde{X} = 12$.

- *n par*

$X_1 = 10 \quad X_2 = 14 \quad X_3 = 12 \quad X_4 = 18 \quad X_5 = 11 \quad X_6 = 23$

Ordenamos los datos:

10 11 12 14 18 23

Posición de la mediana $\Rightarrow \frac{6+1}{2} = 3.5$

Obtenemos la mediana promediando el tercer y cuarto dato: $\tilde{X} = \frac{12+14}{2} = 13$.

Notar que (n+1)/2 no es la mediana, sino la localización de la mediana en el conjunto ordenado de datos.

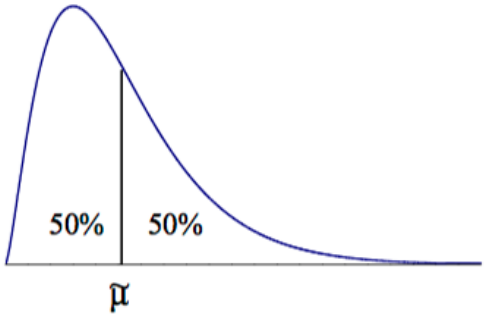
Si hay datos repetidos deben ser incluidos en el ordenamiento.

La mediana es muy simple de obtener a partir de un gráfico de tallo-hojas.

MEDIANA POBLACIONAL

La mediana poblacional se define de modo equivalente a la mediana muestral y es el valor de la variable por debajo del cual se encuentra a lo sumo el 50% de la población y por encima del cual se encuentra a lo sumo el 50% de la población.

La denotamos como $\tilde{\mu}$.



Propiedades de la mediana:

a) La mediana puede ser usada no solo para datos numéricos sino además para datos ordinales, ya que para calcularla solo es necesario establecer un orden en los datos.

b) Si la distribución de los datos es aproximadamente simétrica la media y la mediana serán aproximadamente iguales.

Si la distribución de los datos es asimétrica, la media y la mediana diferirán según el siguiente patrón:

Asimetría derecha (cola larga hacia la derecha) $\Rightarrow$ $\bar{X} > \tilde{X}$

Asimetría izquierda (cola larga hacia la izquierda) $\Rightarrow$ $\bar{X} < \tilde{X}$

Ejemplos:

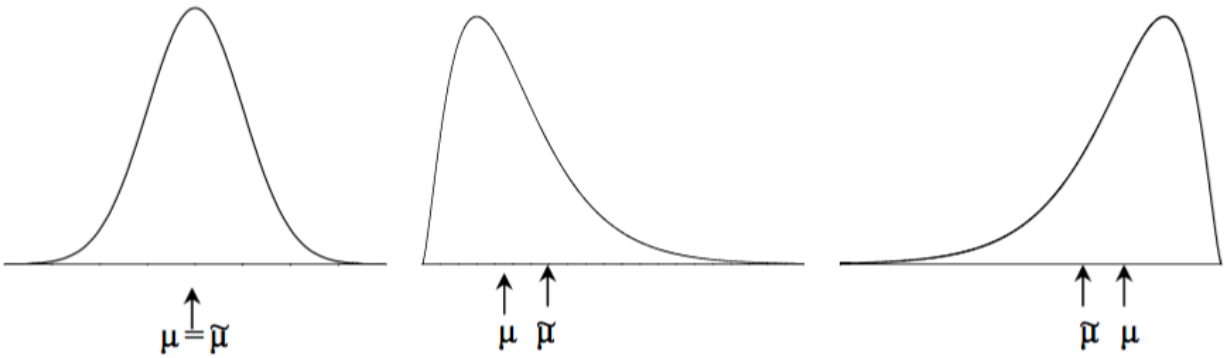
1) 12, 13, 14, 15, 16 $\quad \bar{X} = \tilde{X} = 14$

2) 12, 13, 14, 15, 20 $\quad \bar{X} = 15 > \tilde{X} = 14$

3) 2, 13, 14, 15, 16 $\quad \bar{X} = 12 < \tilde{X} = 14$

En la población:

c) La mediana es una medida de posición robusta. No se afecta por la presencia de datos outliers, salvo que modifiquemos casi el 50% de los datos menores o mayores de la muestra (la proporción de datos que debemos modificar para modificar la mediana depende del número de datos de la muestra).



Ejemplo:

I) 10 11 12 12 13 14 $\quad \bar{X} = 12 \quad \tilde{X} = 12$

II) 10 11 12 12 13 26 $\quad \bar{X} = 14 \quad \tilde{X} = 12$

d) La mediana es insensible a la distancia de las observaciones al centro, ya que solamente depende del orden de los datos. Esta característica que la hace robusta, es una desventaja de la mediana.

Ejemplo. Todos los conjuntos de datos siguientes tienen mediana 12:

I)	10	11	12	13	14
II)	10	11	12	13	100
III)	0	11	12	12	12
IV)	10	11	12	100	100

e) Si hay datos censurados eventualmente puede calcularse la mediana, en la muestra no es posible calcular la media, sin embargo, eventualmente puede calcularse la mediana.

Tiempo de supervivencia (en meses) de pacientes con cierta patología. Los datos que se indican entre paréntesis tienen censura derecha, es decir, se sabe que el paciente sobrevivió ese tiempo, pero no se conoce el tiempo real de supervivencia.

I) 1 5 10 12 18 24 25 28 39 45 (45) 48 50 51 (84) $n = 15$

Como $n = 15$ la mediana es el octavo dato, por lo tanto $\tilde{X} = 28$. Es posible calcularla aunque haya datos censurados, porque los mismos se encuentran más allá de la posición 8 que define la mediana. Aunque no conocemos exactamente el tiempo que sobrevivió el paciente cuyo dato es (45) sabemos que en esta muestra ese dato ocupará la posición 11 o una superior.

II) 1 5 10 (12) 18 24 25 28 39 45 (45) 48 50 51 (84) $n = 15$

No es posible calcular la mediana debido al dato indicado como (12). Sabemos que este paciente sobrevivió por lo menos 12 meses, pero desconocemos el verdadero valor, el que puede ocupar cualquier posición entre la quinta y la última. Comparación de la media y la mediana

	MEDIA	MEDIANA
VENTAJAS	Usa toda la información que proveen los datos. Es de manejo algebraico simple.	Representa el centro de la distribución (en un sentido claramente definido). Robusta a la presencia de outliers. Útil para datos ordinales.
DESVEN-TAJAS	Muy sensible a la presencia de datos outliers.	Usa muy poca información de los datos.

LA MEDIA α-PODADA

La media α-podada es un compromiso entre las dos medidas de posición presentadas. Es una medida más robusta que la media, pero que usa más información que la mediana. La media α-podada se calcula despreciando $n.\alpha$ datos de cada extremo y promediando las observaciones centrales del conjunto ordenado de datos.

¿Cómo calculamos la media α-podada de una muestra de n observaciones?

1. Ordenamos los datos de menor a mayor.
2. Excluimos los $n.\alpha$ datos más pequeños y los $n.\alpha$ datos más grandes.
3. Calculamos el promedio de los datos restantes y lo denominamos $\overline{X}_\alpha$.

¿Cómo elegimos α?

Depende de cuantos outliers se pretende excluir y de cuán robusta queremos que sea la medida de posición.

Cuando seleccionamos $\alpha = 0$ tenemos la media, si elegimos el máximo valor posible para α (lo más cercano posible a 0.5) tenemos la mediana. Cualquier poda intermedia representa un compromiso entre ambas.

Una elección bastante común es α = 0.10, que excluye un 20% de los datos. ¿Cuándo usar esta medida?

Cuando se sospecha que hay errores groseros en los datos, pero no tenemos modo de decidir si el dato es erróneo.

Esto permite excluir datos aberrantes de un modo menos sesgado, porque estamos excluyendo datos de ambos extremos.

Ejemplo:

Calculamos la media 20% podada para los datos siguientes que corresponden a los puntajes asignados a una gimnasta por 5 jueces durante una competencia olímpica.

$$X_1 = 85 \qquad X_2 = 98 \qquad X_3 = 99 \qquad X_4 = 95 \qquad X_5 = 98$$

1. Ordenamos los datos: 85 95 98 98 99
2. Calculamos el número de datos que podaremos en cada extremo

$$n \cdot \alpha = 5 \cdot 0.20 = 1$$

Excluimos el primer y el último dato de la muestra ordenada.

3. Promediamos los datos restantes

$$\overline{X}_{0.20} = \frac{95+98+98}{3} = 97.$$

Para estos datos el promedio y la mediana resulta ser $\overline{X} = 95$, $\tilde{X} = 98$.

¿Qué ventaja tiene haber usado la media 20% podada?

El puntaje final de la gimnasta no se ve afectado por la calificación notablemente baja que le asignara uno de los jueces.

¿Qué hacer cuando el número de datos que debe excluirse no es entero?

Si n = 37 y quisiéramos una poda del 10% deberíamos excluir 37 · 0.10 = 3.7 datos de cada extremo.
Las opciones son:

Seleccionar una poda menor o igual que α. En este caso podamos 3 datos de cada extremo e informamos que se calculó la media 8.1% podada.

Calculamos la media podando 3 datos y luego la media podando 4 datos de cada extremo y finalmente calculamos un promedio ponderado de estas dos medidas.

¿Cuál de las tres medidas de posición preferir: media, mediana o media α-podada?

Si la distribución de la variable es simétrica las tres medidas deberían dar resultados similares. En este caso, es preferible usar la media ya que es la que tiene menor error de estimación. Esto es, la distancia entre la media muestral y la verdadera media poblacional en promedio es menor que la distancia entre la mediana o la media α-podada y la media poblacional.

Si la distribución es asimétrica o con outliers generalmente es preferible resumir los datos con la mediana o la media α-podada, ya que la estimación obtenida en una muestra en promedio se encuentra más cercana al correspondiente parámetro (media poblacional y mediana poblacional).

LA MODA

La moda es el dato que ocurre con mayor frecuencia en el conjunto.

Es una medida de poca utilidad salvo para datos categóricos en los que suele interesar identificar la categoría con mayor cantidad de datos. En una muestra de datos numéricos, puede ocurrir que la moda sea un valor que se repite un cierto número de veces, pero que no es típico. Cuando se considera la distribución poblacional de una variable continua, decimos que esta es UNIMODAL si presenta un pico y BIMODAL si aparecen dos picos claros.

CUARTILES Y OTROS PERCENTILES

Los percentiles son otro modo de resumir una distribución muestral o poblacional.

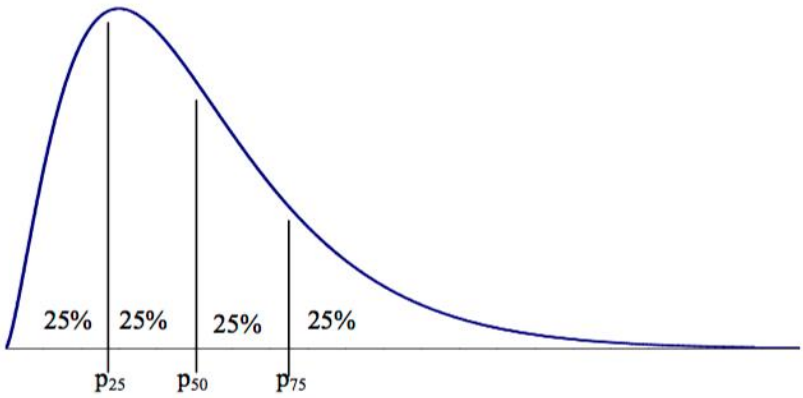
> El *percentil p%* de un conjunto de datos es la observación que deja a lo sumo p% de las observaciones debajo de él y a lo sumo (1 – p)% encima de él.

Como ejemplo, consideremos la distribución de peso de recién nacidos de sexo femenino y 38 semanas de gestación.

Si se informa que el percentil 10% de esta distribución es 2450 g y el percentil 90% es 3370g, estamos indicando que un 10% de las niñas que nacen en la semana 38 de gestación pesan 2450 g o menos (y en consecuencia, 90% pesan más que 2450 g) y que el 90% de las niñas de esta edad gestacional nacen con peso menor o igual que 3370 g (y sólo el 10% con peso mayor que 3370 g).

La mediana es el percentil 50%.

Otros percentiles con nombre propio son el percentil 25% y el percentil 75% que se denominan cuartil inferior y superior respectivamente, ya que juntamente con la mediana dividen a la distribución en 4 porciones iguales.



¿Cómo se calculan los cuartiles de una muestra de n observaciones?

1. Ordenar los datos de menor a mayor.
2. El cuartil inferior es el dato que ocupa la posición $(n+1)/4$ en la muestra ordenada.
3. El cuartil superior es el dato que ocupa la posición $3(n+1)/4$ en la muestra ordenada.

   Si la posición resulta ser un número decimal, promediamos los datos que se encuentran a izquierda y derecha de la posición obtenida.

Ejemplo:

Consideremos los siguientes datos ordenados (n = 13).

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13
Datos	104	112	134	146	155	168	170	195	246	302	338	412	678

$$\text{Posición del Cuartil Inferior} = (13+1)/4 = 3.5 \quad \Rightarrow C_I = \frac{134+146}{2} = 140$$

$$\text{Posición de la mediana} = (13+1)/2 = 7 \quad \Rightarrow \tilde{X} = 170$$

$$\text{Posición del Cuartil Superior} = 3.(13+1)/4 = 10.5 \quad \Rightarrow C_S = \frac{302+338}{2} = 320$$

CINCO NÚMEROS RESÚMENES

Un modo de resumir toda la distribución de los datos es informar los siguientes cinco números resúmenes:

Mínimo, Cuartil inferior, Mediana, Cuartil superior, Máximo.

En nuestro ejemplo:

Mínimo =	104	25%
Cuartil Inferior =	140	25%
Mediana =	170	25%
Cuartil Superior =	320	25%
Máximo =	678	

MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Las medidas de posición dan una idea de dónde se encuentra el centro de la distribución, pero no nos dicen cuán disperso es conjuntos de datos:

Muestra A:	55	55	55	55	55	55	55
Muestra B:	47	51	53	55	57	59	63
Muestra C:	39	47	53	55	57	63	71

En todos ellos $\bar{X} = \tilde{X} = 55$, pero las muestras difieren notablemente.

Las medidas de dispersión o variabilidad describen cuán cercanos se encuentran los datos entre ellos, o cuán cerca se encuentran de alguna medida de posición. Introduciremos a continuación algunos estadísticos que miden variabilidad del conjunto de datos.

RANGO MUESTRAL

El *rango* de *n* observaciones X_1, X_2, ..., X_n es la diferencia entre la observación más grande y la más pequeña

$$\text{Rango} = \text{máx}(X_i) - \text{mín}(X_i)$$

Ejemplo:

Muestra A:	55	55	55	55	55	55	55	Rango = 55 – 55 = 0
Muestra B:	47	51	53	55	57	59	63	Rango = 63 – 47 = 16
Muestra C:	39	47	53	55	57	63	71	Rango = 71 – 39 = 32

Características y propiedades:

- ✓ Es muy simple de obtener.

- ✓ Es extremadamente sensible a la presencia de datos atípicos. Si hay datos outliers, estos estarán en los extremos, que son los datos que se usan para calcular el rango.
- ✓ Ignora la mayoría de los datos.
- ✓ En general aumenta cuando aumenta el tamaño de la muestra (las observaciones atípicas tienen más chance de aparecer en una muestra con muchas observaciones). En consecuencia, reportar el rango o el máximo y el mínimo de un conjunto de datos, no informa demasiado sobre las características de los datos. A pesar de esto es frecuente encontrar en las publicaciones científicas datos numéricos resumidos a través de una medida de posición acompañada por los valores mínimo y máximo.

DESVIACIÓN ESTÁNDAR Y VARIANZA MUESTRAL

La desviación estándar mide cuán lejos se encuentran los datos de la media muestral. Un modo de medir la variabilidad de los datos de una muestra sería tomar algún valor central, por ejemplo, la media, y calcular el promedio de las distancias a ella. Mientras mayor sea este promedio, más dispersión deberían presentar los datos. Sin embargo, esta idea no resulta útil, ya que las observaciones que se encuentran a la derecha de la media tendrán distancias (o desviaciones) positivas, en tanto que las observaciones menores que la media tendrán distancias negativas y la suma de las distancias a la media será inevitablemente igual a cero. Un modo de evitar este inconveniente es elevar las distancias al cuadrado y de este modo tener todos sumandos positivos.

Definimos la varianza de una muestra de observaciones X_1, X_2, ..., X_n, cuya media es $\overline{X}$, como

$$s^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

La varianza muestral puede pensarse como "promedio" de las distancias a la media al cuadrado. Sin embargo, la varianza no tiene las mismas unidades que los datos. Para salvar este inconveniente, definimos la *desviación estándar muestral* como la raíz cuadrada positiva de la varianza:

$$s = \sqrt{s^2}$$

VARIANZA Y DESVIACIÓN ESTÁNDAR POBLACIONAL

Si se dispone de la información de una variable X para las N unidades de análisis de la población, denotamos con σ^2 y σ (sigma) la *varianza* y la desviación estándar de la población respectivamente y las definimos del siguiente modo:

$$\sigma^2 = \frac{(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad \sigma = \sqrt{\sigma^2}$$

La razón para usar (n – 1) y no n en el denominador de la varianza muestral tiene que ver con el hecho de que el valor de s^2 obtenido en una muestra, se usa para estimar la varianza poblacional σ^2. Definida con (n – 1) en el denominador la varianza muestral posee una propiedad deseable, resulta ser insesgado, esto es, en promedio no subestima ni sobrestima el valor de la varianza poblacional.

Ejemplo:

Muestra A:	55	55	55	55	55	55	55	$s^2 = 0$	$s_A = 0$
Muestra B:	47	51	53	55	57	59	63	$s^2 = 28$	$s_B = 5.29$
Muestra C:	39	47	53	55	57	63	71	$s^2 = 108$	$s_C = 10.39$

Calculamos la varianza y el desvío estándar para la Muestra B. Se deja como ejercicio verificar que los resultados obtenidos para A y C son correctos.

$$s_B^2 = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} = \frac{(47-55)^2 + (51-55)^2 + \cdots + (63-55)^2}{7-1}$$

$$= \frac{(-8)^2 + (-4)^2 + \ldots + 8^2}{6} = \frac{168}{6} = 28$$

$$s_B = \sqrt{28} = 5.29$$

Comparando las desviaciones estándar de las tres muestras vemos que $s_A < s_B < s_C$. Además, observamos que $s_A = 0$, ya que todas las observaciones toman el mismo valor.

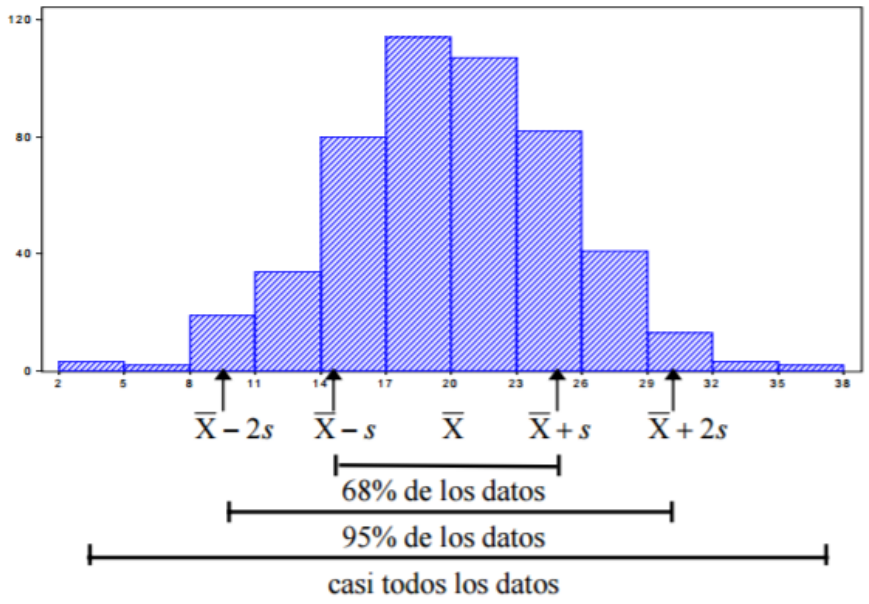
Interpretación del valor de la desviación estándar

La desviación estándar *s* es útil para comparar la variabilidad de dos conjuntos de datos en los que la variable ha sido medida en las mismas unidades. Si en una muestra *s* = 5.4 y en otras = 10.4 podemos asegurar que los datos de la segunda muestra están más dispersos que los de la primera. Pero ¿cómo interpretamos el valor *s* = 5.4? La desviación estándar nos da idea de la distancia promedio de los datos a la media (aunque estrictamente hablando no es el promedio). Pero la interpretación de s requiere algún conocimiento de la distribución de los datos.

Regla empírica

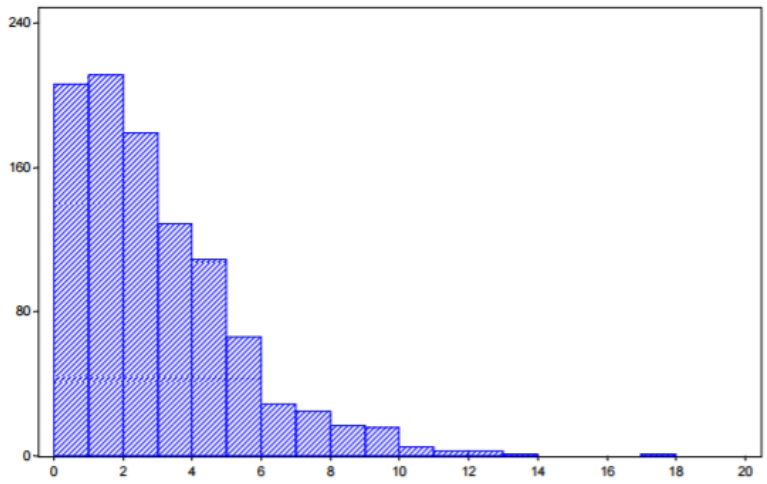
- Aproximadamente el 68% de las observaciones caen en el intervalo $\overline{X} - s$ y $\overline{X} + s$.
- Aproximadamente el 95% de las observaciones caen en el intervalo $\overline{X} - 2s$ y $\overline{X} + 2s$.
- Prácticamente todas las observaciones caen en el intervalo $\overline{X} - 3s$ y $\overline{X} + 3s$.

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces,



Esta regla es válida para distribuciones no necesariamente acampanadas, pero puede ser errónea cuando se aplica a distribuciones fuertemente asimétricas tales como la que se presenta en el histograma siguiente en el que $\overline{X} = 3$ y $s = 2.45$.

Esta distribución ficticia podría representar la distribución de ingreso mensual (en cientos de pesos) de una muestra de asalariados con cargos no jerárquicos de una provincia argentina (por ejemplo).



¿Es útil nuestra regla empírica para el desvío estándar en datos con esta distribución? En este caso, al restar 2s a la media, caemos fuera de la escala de la variable $\overline{X} - 2s = 3 - 2 \cdot 2.45 = -1.9$ y la interpretación que propusimos a través de la regla empírica resulta no ser apropiada. Cuando la variable sólo puede tomar valores dentro de un cierto rango, tal como ocurre con el ingreso o el tiempo transcurrido hasta un cierto evento que no pueden ser menores que cero, el hecho de obtener valores fuera del rango al aplicar la regla con 1 o 2 desvíos estándar nos indica que la distribución de la variable es fuertemente asimétrica.

Propiedades de la desviación estándar:

- *s* mide la dispersión alrededor de la media, por lo tanto, es natural elegir esta medida de dispersión cuando se usa la media como medida de posición.
- $s = 0$ solamente cuando todos los datos son iguales, de otro modo $s > 0$.
- *s* es una medida de dispersión muy sensible a la presencia de datos outliers. De hecho, es más sensible que la media ya que las distancias están elevadas al cuadrado.

En la siguiente unidad se te presentaran dos medidas de dispersión robustas…

INFORMACIÓN (INCLUÍDA EN ESTE DOCUMENTO EDUCATIVO) TOMADA DE:

Sitios web:

1. http://searchdatacenter.techtarget.com/es/definicion/Analisis-estadistico
2. http://www.ugr.es/~batanero/pages/ARTICULOS/trabajomasterPedro.pdf
3. http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf